

ORIGINAL RESEARCH

Open access

Leveraging Chain-of-Thought Prompt Engineering to Enhance Knowledge Creation through Improvable Student Ideas

Laura Keller^{1*}, Thomas Lehmann¹, Simon Brunner², Christoph Meier³, David Thompson⁴

Abstract

Within a student learning community, the generation of new knowledge represents a vital mechanism for expanding collective understanding and driving original synthesis. Students assume critical responsibilities when they target knowledge deficiencies and engage in cooperative efforts to refine communal concepts emerging from dialogue. Nonetheless, the standard of these concepts can frequently be constrained by resource deficits or a narrow spectrum of viewpoints. The introduction of generative Artificial Intelligence and large language models (LLMs) into educational settings has unlocked new opportunities for analyzing and enriching idea-focused debates that were formerly unachievable. Maximizing the utility of LLMs necessitates specialized competencies in prompt engineering, particularly regarding the in-context methodology termed Chain-of-Thought (CoT) to formulate and scaffold progressive concepts within student conversations. This investigation gathered a dataset of 721 conversational entries—encompassing 272 applicable question–answer dynamics and 149 distinct discussion tracks—produced by 31 participants during a 2-day student Knowledge Building Design Studio (sKBDS). The researchers augmented student inputs utilizing the CoT technique and subsequently evaluated the AI-generated outputs against the original student assertions. The outcomes are demonstrated through two explicit conversational tracks, confirming that CoT-configured prompts enable LLMs to output responses that actively advance malleable concepts during knowledge construction. This inquiry examines data from genuine classroom discourse and offers actionable insights for both educational research and instructional implementation.

Keywords Large language models, Prompt engineering, Knowledge creation, Knowledge building, Chain-of-thought

*Correspondence:

Laura Keller

laura.keller@usys.ethz.ch

¹ Department of Prompt Engineering and Educational Technology, Faculty of Education, ETH Zurich, Zurich, Switzerland

² Department of Chain-of-Thought and Knowledge Creation, Faculty of Education, EPFL Lausanne, Lausanne, Switzerland

³ Department of Improvable Ideas and Student Learning, Faculty of Education, University of Bern, Bern, Switzerland

⁴ Department of AI in Education and Prompt Design, Faculty of Education, University of Basel, Basel, Switzerland

Introduction

The adoption of standard large language models (LLMs) has sparked profound academic interest across a broad range of scientific areas and operational domains, demonstrating capacities that span from resolving elementary inquiries to unpacking intricate intellectual dilemmas and writing functional source code [1]. In educational research, the creation of knowledge is recognized as a vital architecture of activities through which a learning cohort broadens its shared intelligence, a process fundamentally linked to innovation [2]. While the capacity of young learners to formulate entirely novel paradigms may not match that of professional researchers, they can be successfully oriented along a knowledge-

building trajectory. Through this, they cultivate parallel competencies, such as pinpointing significant gaps in the group's current understanding, establishing strategic objectives, and identifying collaborative methods to progressively advance and elevate community concepts over time [3].

In recent years, modern technological innovations have aided learners in sustaining these knowledge-building conversations via web-based interactive environments [4] (specifically Knowledge Forum). This system is custom-designed for knowledge creation and construction, offering an environment where students co-develop conceptual frameworks while simultaneously acting as a permanent digital repository of the group's intellectual progression over time.

Although the contributions of students toward generating knowledge have been extensively investigated and mapped alongside deeper insights into behavioral responses [5, 6], the qualitative standard of these interactions remains a complex issue. It is contingent upon a diverse array of variables, including internal variables like personal drive, alongside the accessibility and heterogeneity of concepts that inspire students to deepen their inquiries through continuous debate. On occasion, student outputs are either inadequate or fall short of the intellectual thresholds that educators aim to establish in specialized instructional environments. The advent of generative AI—particularly through the widespread availability of Generative Pre-trained Transformer (GPT) frameworks like ChatGPT—introduces an innovative technological resource. This tool empowers both instructors and learners to interrogate and unpack problems and concepts in manners that were completely unavailable to previous generations.

Even so, navigating and communicating efficiently with LLMs requires individuals to develop competence in prompt engineering, which centers on constructing explicit directives for the AI to tailor its responses and maximize its operational capacity. The primary objective of prompt engineering involves the systematic design of instructional inputs deployed for subsequent operations, thereby guiding pre-trained frameworks to execute target operations and hit explicit performance benchmarks [7]. Utilizing LLMs for highly complex operations—including mathematical reasoning, everyday practical logic, and symbolic manipulation—has proven exceptionally difficult [8], particularly when the objective centers on executing tasks anchored in common-sense deductions. Methodologies designed to optimize model outputs generally encompass either the structural fine-tuning of pre-trained architectures or the application of in-context prompt designs to steer the models toward specific task fulfillment.

Rather than pursuing the conventional methodology of fine-tuning LLMs through gradient updates [9]—a strategy that is both financially demanding and structurally laborious—an expanding cohort of investigators favors the deployment of in-context prompting with LLMs. This technique leverages a limited series of input–output examples to secure heightened operational accuracy [10]. This practice of utilizing explicit instructions to direct model execution [11], frequently classified as prompt learning, empowers language architectures to achieve few-shot or zero-shot execution, adapting smoothly to unfamiliar settings with a minimal requirement for annotated data [12]. The present inquiry isolates and analyzes the application of prompt engineering within knowledge-generating environments, structured around the following central research question: “How can the approach of Chain-of-Thought [13] (CoT) be applied to support students' improvable ideas in a knowledge creation environment?”

It is hypothesized that integrating the Chain-of-Thought (CoT) framework within prompt designs will yield more comprehensive and profound mechanisms for anchoring student concepts in a knowledge-generative setting. Consequently, this intervention is expected to expand the students' capacity for formulating and refining concepts, while concurrently elevating the descriptive caliber of the collective dialogue supported by LLM applications. The significance of this architectural shift resides in how a prompting-centric methodology can expose expanded horizons for AI integration without requiring massive foundational training datasets, preserved alongside the capacity of a static, single-checkpoint model to execute varied tasks without relinquishing its broader contextual utility. By deploying Chain-of-Thought prompting strategies for knowledge construction and practical deductive reasoning, this investigation aims to enrich the broader domain of applied AI in educational systems.

Background and Literature

Emergence of generative AI and LLMs

Generative AI frameworks have existed within computing paradigms since the 1960s in the form of foundational conversational agents, though they have experienced a dramatic surge in prominence over the last decade following the introduction of generative adversarial networks (GANs) in 2014 by Goodfellow and his peers [14]. GANs were quickly recognized as a highly disruptive technological leap, given their remarkable capacity for replicating complex data distributions, generating original assets, and rendering authentic representations across diverse fields of study. Yet, alongside these manifest structural benefits and broader societal influences, negative implications also emerged, specifically regarding the distribution of deceptive media formations such as deepfake recordings. These assets effectively obscured the boundaries separating empirical historical recordings from artificially synthesized media, severely complicating verification metrics [15].

These manifestations of information distortion do not diminish the substantial advantages generated by an alternative lineage of neural network models predicated on the transformer architecture, specifically designated as the Generative Pre-trained Transformer (GPT) [10]. This landmark breakthrough in computational intelligence underlies consumer systems like ChatGPT, which achieved widespread public adoption in late 2022, presenting a significantly more robust and approachable medium capable of driving future scientific exploration toward artificial general intelligence and systems engineered to resolve human-tier challenges. Within educational frameworks, this technology has already been integrated into instructional and learning workflows, bringing distinct advantages and vulnerabilities [16, 17], pedagogical architectures and implementations [18], and systematic approaches [19] oriented around the ethical deployment of ChatGPT—a foundational element required to foster expanded acceptance among instructional staff and student populations.

Role and various approaches of prompt engineering

Prompt engineering, catalyzed by the mechanics of in-context adaptation and instructional prompting, assumed central importance following the mainstream adoption of ChatGPT, as the fundamental mode of communication and control with an LLM involves submitting textual statements to elicit a corresponding output—essentially serving as a operational directive to the underlying model. In executing communication loops with an LLM like ChatGPT, prompts—comprising queries, assertions, and declarations—are submitted with the expectation of obtaining a highly precise and sufficient resolution.

Mirroring the communication patterns found in real-world human conversations, these structural prompts can exhibit highly divergent parameters of transparency and definition. Fundamentally, optimized instructions capable of generating high-tier responses from LLM architectures must satisfy two primary benchmarks: clarity and specificity [11]. Clear instructions utilize uncomplicating terms and direct, linear syntax that eliminates ambiguity and avoids unnecessary technical jargon. Conversely, prompts optimized for specificity mandate the integration of all required contextual boundaries so that the LLM can accurately resolve the question or execute the assigned task.

Instructional prompt formulation also manifests across several operational modalities. Within these varied strategies, standard prompt construction is viewed as the baseline format, frequently deployed to establish performance benchmarks for LLM systems. Multiple distinct methodologies exist, encompassing the following variations:

- **Zero-shot.** This process involves delivering nothing more than a natural language statement of the target objective to the underlying architecture, omitting any practical examples, and calling directly for a final computation;
- **One-shot.** A solitary contextual example is delivered to the system to guide it toward generating an output aligned with the user's intent, an approach that consistently yields superior accuracy compared to configurations lacking baseline examples;

- **Few-shot.** A curated selection of high-quality examples or demonstrations is embedded into the model's immediate context window at the point of inference. This selection contains matched input–output vectors tailored to the target task, designed fundamentally to steer the system toward outputting results that align closely with human parameters and targets, yielding performance matrices that regularly surpass both zero-shot and one-shot configurations.

Table 1 outlines several operational instances displaying the appearance of prompt formulations when students address thematic topics using zero-shot, one-shot, or few-shot methodologies.

Table 1. Examples of zero-shot, one-shot, and few-shot prompting in question-and-answer format based on topics that students ask.

Approach	Format	Input	Output
Zero-shot	<Question> ?	Answer True or False: Question: Will recycling alone reverse global warming?	False
One-shot	<Question> ? <Answer> <Question> ?	Answer which appliance: Question: Which appliance consumes more electricity in the same period of time, laptop or desktop? Answer: Desktop Question: Which appliance consumes more electricity in the same period of time, water heater or fridge?	Water heater
Few-shot	<Question> ? <Answer> <Question> ? <Answer> <Question> ? <Answer> <Question> ?	Answer good or bad idea: Question: Can we keep appliances on standby when not using them? Answer: Bad idea. Question: Maybe we should use less plastics if they become trash in the ocean? Answer: Good idea. Question: Should all renewable energy sources be solar? Answer: Bad idea. Question: Is it better to encourage eating more vegetables to reduce greenhouse gas emissions?	Good idea.

It is important to note that the application of a strict question-and-answer arrangement is not mandatory; the structural design of the prompt is governed entirely by the requirements of the task. For example, these can include organizational sorting tasks where the system is given contextual instances illustrating a specific sorting process. The concluding entry detailed in **Table 1** under the few-shot demonstration confirms that conceptual frameworks can be comprehensively categorized by the LLM.

Moving beyond foundational in-context prompt applications, supplementary variations of LLM instructional interactions are gaining traction, including fine-tuning methodologies [20], automated script generations [21, 22], and the targeted tailoring of synthesized results. Furthermore, formal repositories have been published [11] to index structured systems for mapping prompt configurations designed to untangle diverse operational obstacles and fit distinct professional fields, detailing patterns that have been successfully deployed to refine the quality of LLM-generated information.

Chain-of-Thought (CoT)

Complex logical processes that frequently encompass intricate concepts and strategies generally necessitate progressive, intermediate phases to achieve alignment and formulate final outcomes—a capability that few-shot prompt methods can demonstrate to a certain degree.

A foundational inquiry [13] introduced a compelling methodology known as Chain-of-Thought (CoT) prompting, which successfully unlocked the latent deductive capacities of large language models by embedding a sequence of intermediate analytical phases. Fundamentally, this instructional strategy prompts the model to execute reasoning patterns that replicate the step-by-step structural logic provided in the prompt itself. This mode of instruction yields several distinct

advantages over traditional prompting techniques: it systematically builds upon and integrates intermediate reasoning steps into active few-shot learning routines, thereby elevating model performance when addressing highly complex problems; it clarifies the interpretation of final outputs and provides deeper insights into the underlying mechanics of model behavior and selection pathways, increasing transparency and comprehension regarding how outputs are generated; and, crucially, it can surpass traditional prompting metrics across tasks requiring mathematical calculation, common-sense deduction, and symbolic manipulation.

Out of these three specific domains (mathematical, common-sense, and symbolic reasoning), and as detailed and illustrated within original CoT literature [13], the investigation detailed in this document concentrates specifically on common-sense deductive processes that correspond with the core tenets of knowledge generation. We mapped these corresponding pathways together to advance the architectural design of prompt engineering within knowledge-generating frameworks.

On one hand, generalized common-sense reasoning—which remains foundational for successfully interacting with the surrounding environment—frequently remains beyond the operational grasp of most standard natural language understanding frameworks [23]. Common-sense deduction centers on resolving inquiries that demand an understanding of routine, everyday scenarios and formulating logical conclusions derived from generalized background knowledge. In the context of knowledge creation, this manifests when students formulate inquiries tied to their authentic lived experiences, extract real-world problems from those occurrences, and subsequently co-develop and refine concepts to resolve those challenges. On the other hand, Table 2 illustrates an instance of common-sense deduction resolved via an LLM, validating this investigation into the deployment of prompt engineering and LLM architectures to support analytical logic and knowledge construction.

Table 2. Example of commonsense reasoning using an LLM.

Question: If John is looking for a location where he can have fun playing in water, which place should he choose? Answer options: (a) an aquarium (b) the ocean (c) a nearby pond (d) a reservoir (e) his bathtub
Answer: (b) the ocean

Several advantageous characteristics for cultivating logical processing within LLMs were isolated in the foundational study by Wei et al. [13], categorized as follows:

1. Models should be permitted to decompose multifaceted issues into a series of intermediate steps, which can draw upon expanded computational resources if necessary;
2. The CoT framework functions as an interpretable window into the internal behaviors of the model, enabling the architecture to evaluate the optimal approach for anchoring an answer;
3. CoT reasoning is applicable to mathematical calculation, common-sense deduction, and symbolic processing, with prospective utility across diverse tasks that humans resolve via linguistic communication;
4. CoT reasoning pathways can be readily extracted from sufficiently scaled, off-the-shelf language models by embedding examples of CoT sequences directly within the contextual demonstrations of few-shot prompts.

As observed with alternative prompting methodologies, CoT instruction can be executed via two primary avenues, mirroring the configurations :

- **Zero-shot CoT.** Natural language declarations are utilized to motivate the model to independently construct analytical logic pathways;

- **Few-shot CoT.** The underlying model is provided with a small selection of high-quality contextual examples or demonstrations to formulate a conclusion. This specific method was implemented in the current investigation to guide the model toward final outcomes, given that the performance advantages are significantly more pronounced when complex reasoning operations are paired with larger language architectures.

By embedding multiple distinct demonstrations of CoT as contextual exemplars within few-shot prompts, researchers can guide LLMs to output corresponding analytical pathways. To illustrate this mechanism, we presented the LLM with an inquiry regarding human environmental sustainability focused on the dense utilization of shared spaces. **Table 3** contrasts the structural appearance of a standard prompt against the delivery of dedicated exemplars within a CoT prompt configuration.

Table 3. Comparing standard prompting against Chain-of-Thought (CoT) prompting.

Approach	Format	Input	Output
Standard prompting	<Question> ? <Answer> <Question> ?	Question: John prefers to go to places that are not very crowded. Where should he go? Answer choices: (a) a shopping mall (b) a park (c) a train station (d) a corridor (e) by the roadside Question: John enjoys visiting highly crowded places. Where should he go? Answer choices: (a) a shopping mall (b) a park (c) a train station (d) a corridor (e) by the roadside	Answer: (b) a park Answer: (a) a shopping mall
CoT prompting (italicized)	<Question> ? <Answer with reasoning> <Question> ?	Question: John prefers to go to places that are not very crowded. Where should he go? Answer choices: (a) a shopping mall (b) a park (c) a train station (d) a corridor (e) by the roadside Question: John enjoys visiting highly crowded places. Where should he go? Answer choices: (a) a shopping mall (b) a park (c) a train station (d) a corridor (e) by the roadside	Answer: The correct choice should be a place with fewer people. From the options, a park tends to have fewer people and is more suitable than a corridor or roadside. Therefore, the answer is (b) a park. Answer: The correct choice should be a place with more people. From the options, a shopping mall, a train station, and sometimes a park during busy hours tend to have more people than a corridor or roadside. So the answer could be either (a) a shopping mall or (c) a train station.

Based on the outputs recorded in **Table 3**, the traditional prompt configuration succeeded in generating a highly plausible solution, although it remains open to debate whether a public transit hub is fundamentally more crowded than a commercial shopping center, which was the solution provided. Standard prompt formulations ensure the generation of a plausible conclusion, yet the internal analytical logic remains hidden from the end user.

In contrast, under the CoT prompt configuration, it is clear that while a direct, unyielding conclusion was feasible, it was bypassed due to the bounding data limitations inherent to in-context prompting. Nevertheless, the primary utility of CoT prompting manifests in the delivery of supplementary analytical parameters that assist the LLM user in determining the most contextually accurate resolution—which, in this specific instance, could be either a commercial shopping center or a

transit station, depending on peak operational hours. This exemplifies how integrating CoT frameworks into prompt designs can markedly expand the capacity of an LLM to navigate complex logical challenges.

As a component of this broader literature assessment, we recognize several emerging innovations designed to further optimize the analytical output of current CoT frameworks. One notable development involves replacing the foundational greedy decoding strategies originally utilized in CoT prompting with a self-consistency mechanism [24]. This approach capitalizes on the human intuition that multiple disparate analytical pathways can converge on a single, unique, and correct conclusion. This innovation operates on the principle that when an issue demands more rigorous deliberation and systematic breakdown, it naturally accommodates a wider diversity of reasoning trajectories that culminate in a unified solution [25]. Consequently, this architectural advancement represents a highly promising avenue for optimizing CoT methods to support the workflows of knowledge creation and collaborative construction. However, we explicitly clarify that this specific framework falls outside the core analytical scope of this paper.

Knowledge creation and knowledge building

Complex analytical operations that frequently involve intricate concepts and blueprints generally require progressive intermediate steps to foster convergence and generate new paradigms of understanding. Before this state can be achieved, educational systems must be structurally realigned to meet the operational demands of knowledge generation within contemporary digital environments. Knowledge building functions similarly to the knowledge creation methods deployed within advanced scientific research laboratories, emphasizing tangible value for the individual learner, the immediate community, and broader society. This approach has evolved alongside theoretical and instructional models over several decades, systematically opening these knowledge-generative practices to school-aged cohorts [26].

Knowledge building, operating as a principle-driven pedagogy [27] and a structured suite of practices, empowers students to interrogate overarching thematic questions (e.g., Does a reduction in single-use plastics make a difference to climate change?) and collectively formulate and advance concepts through epistemic agency, improvable discourse, and ideas. By positioning student concepts within a shared, public environment—such as the Knowledge Forum [4]—learners can visually map and explore highly diverse perspectives and conceptual frameworks, building upon one another's ideas as a collective community endeavor to expand their shared intelligence. This sustained design implementation research represents a highly systematic methodology capable of restructuring educational practices to establish self-improving frameworks for ongoing integration within knowledge-generative domains, carrying profound implications for nurturing self-correcting networks and cohorts that leverage technical systems to realign contemporary education with active knowledge creation [26].

Materials and Methods

CoT design for knowledge creation

In traditional educational environments, learners typically build on peer concepts or draw insights from external publications; however, the integration of generative AI through large language models offers a novel, dynamic informational stream to facilitate the modification and advancement of student ideas. By leveraging the CoT prompt framework outlined in and filtering it through a knowledge-generative perspective designed to cultivate malleable ideas within a collaborative learning environment, researchers can implement the following systematic blueprint to architect CoT models for knowledge creation:

1. Student inquiries generated during collaborative learning can be systematically deconstructed into multi-stage logical progressions to serve as target input prompts;
2. The foundational context of the CoT prompt can be enriched with detailed baseline variables to establish a robust interpretative scope, thereby granting sufficient latitude for the generated text to fully articulate and expand upon the core concepts;

- 3. CoT reasoning mechanisms can be pointedly directed toward everyday logic tasks, matching the operational requirements necessary for constructing new understanding within collaborative spaces;
- 4. Structured reasoning pathways can be unlocked within the language model by integrating explicit step-by-step cognitive trajectories directly into few-shot context configurations.

Table 4 highlights several operational models optimized for executing few-shot CoT prompting within knowledge-generative frameworks, illustrating the baseline designs used to extract step-by-step analytical outputs from the LLM.

Table 4. Few-shot exemplars showing how we can add Chain-of-Thought prompting (italicized) to existing questions and responses for knowledge creation.

Question 1: If the goal is to lower carbon emissions, what activity should be discontinued? Answer options: (a) using natural gas (b) using fossil fuels (c) using wind energy (d) using solar energy Answer 1: The response should identify the option that most contributes to carbon emissions reduction. Among the given choices, fossil fuels are the primary source of high carbon emissions. Therefore, the correct answer is (b) using fossil fuels. —Separate question-and-answer— Question 2: Which commuting method to school is the most environmentally friendly and sustainable? Answer options: (a) by private electric car (b) by public diesel bus (c) by public electric taxi (d) by public electric trains Answer 2: The answer should consider both sustainability and the efficiency of public transportation compared to private transport. Among the options, electric-powered public transport is the most environmentally friendly and sustainable. Hence, the correct answer is (d) by public electric trains.

This research explored how CoT analytical pathways can be combined with common-sense processing to drive knowledge creation. This was accomplished by transforming authentic student inquiries gathered from actual classrooms into structured CoT prompts, running them through an LLM, and evaluating the resulting text against the baseline student dialogue to determine the overall utility of the prompt configuration. The resulting structural comparisons and output profiles offer actionable frameworks for both instructors and students, guiding them on how to leverage CoT architectures to anchor their knowledge-building milestones and accelerate student-driven conceptual refinement through LLM tools.

Settings, participants, and setup

Context and nature of student questions

This inquiry utilized a quasi-experimental research framework, harvesting empirical data from a live educational setting configured as a student Knowledge Building Design Studio (sKBDS). The central investigative focus revolved around ecological sustainability, and the event transpired in November 2021. Prior to data collection and subsequent discourse analysis, institutional ethics approval was sought and formal clearance was granted by the university. A cohort of 31 student participants spanning primary and secondary educational levels (Grades 5 through 10) participated in a 2-day, web-facilitated investigation of real-world environmental crises linked to the core theme, logging their peer-to-peer dialogues on the Knowledge Forum digital platform. The qualitative nature of these student inquiries represented a cross-section of natural and social sciences, ranging from elementary requests for definitions to complex problem statements containing explicit hypotheses and technical explanations as the participants worked to resolve their questions.

Extraction of question–answer pairs

The paired question–answer interactions logged by participants on the Knowledge Forum consisted of conversational sequences where students actively interrogated and answered thematic issues. These clusters were predominantly used to secure clarification on a debated point, evaluate a recently concluded learning task, or organize conceptual outlines for group presentations. A complete, valid question–answer pair requires an inquiry that is explicitly situated within the sKBDS environmental sustainability domain (e.g., “How did carbon come about?”) rather than falling into strictly informal or casual socializing (e.g., “What drinks? Bubble tea?”), followed by an appropriate, substantive reply from a peer within the sKBDS framework. The temporal distribution of these responses was asynchronous and limited to data generated across the 2 days of active studio facilitation.

As a core element of the data-cleaning pipeline, and to satisfy the explicit analytical parameters of this study, only complete, matched question–answer pairs were retained for final evaluation. This filter was required to accurately weigh the synthesized LLM outputs against genuine human responses, allowing for a structured critique of CoT functional utility. From the baseline pool of 721 conversational turns logged during the 2-day sKBDS session, 272 complete, contextually aligned question–answer pairs were isolated as viable data points, clusters which collectively generated 149 distinct dialogue threads. Within this document, we highlight two specific complete question–answer dialogue tracks in the empirical results to illustrate the latent capacity of CoT setups to nurture progressive concepts within knowledge-generative and collaborative learning environments.

Extraction of question–answer pairs

Through the execution of this study, we sought to systematically clarify the performance value of CoT by isolating complete scientific and social-scientific question–answer pairings and reinforcing them with contextual CoT few-shot demonstrations before processing them as inputs into the language model. As visualized in the operational archetypes in **Tables 3 and 4**, the proposed structural modifications allow AI-generated responses to be evaluated directly against baseline human feedback. Mechanistically, the methodology dictates that two or more question–answer pairings must occur sequentially and maintain mutual relevance to permit legitimate comparison; furthermore, successive replies do not need to originate from the identical human respondent, given that the underlying objective centers on measuring response quality and identifying divergence vectors between an AI-synthesized response and the original human entry. **Figure 1** outlines the operational pathway through which these responses can be mapped against one another. The Results section documents the explicit differences and qualitative comparisons recorded between the LLM-derived information and the students' baseline inputs.

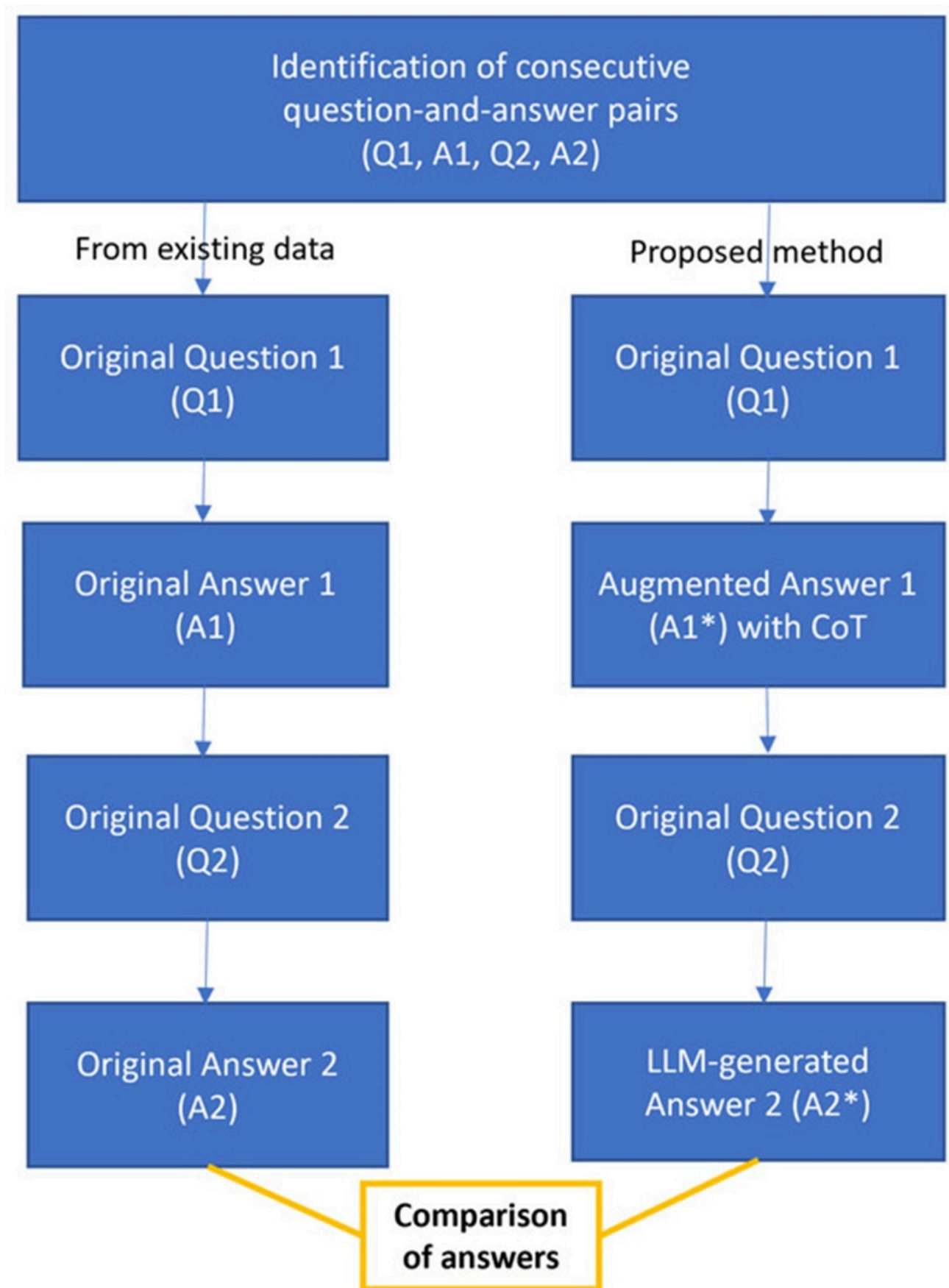


Figure 1. Comparison of answers A2 and A2*, provided respectively by humans (left track) and generated by the LLM using augmented exemplars with CoT few-shot prompting (right track).

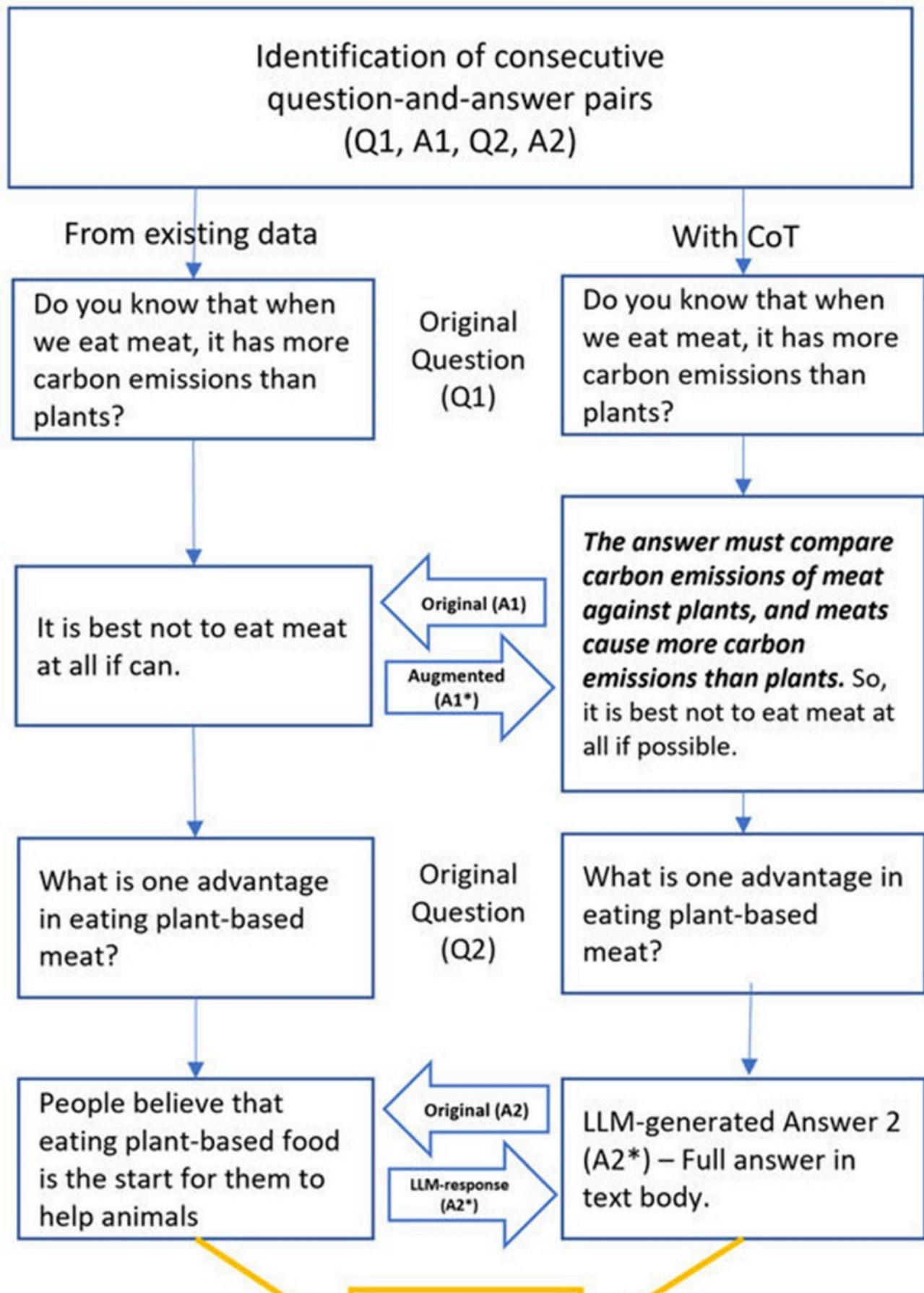
Choice of Large Language Model (LLM)

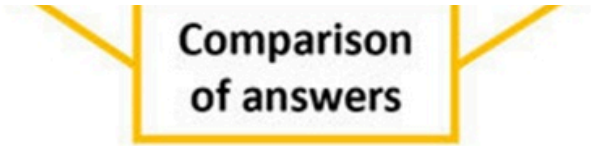
The GPT-3.5 Turbo architecture was selected as the operational large language model for this investigation. While sharing baseline characteristics with contemporary language networks like BERT and PaLM, GPT's robust performance standard across a broad spectrum of computational tasks and disciplinary subjects positioned it as the optimal selection. Furthermore, the straightforward accessibility of GPT-3.5 Turbo through an application programming interface (API) represented an additional motivating factor for its operational deployment. Preliminary benchmarking using emerging systems (such as GPT-4 and GPT-4o) yielded response qualities matching our analytical projections, indicating that the baseline framework proposed in this study is positioned to perform efficiently within knowledge-generative tasks across upcoming iterations of large language models.

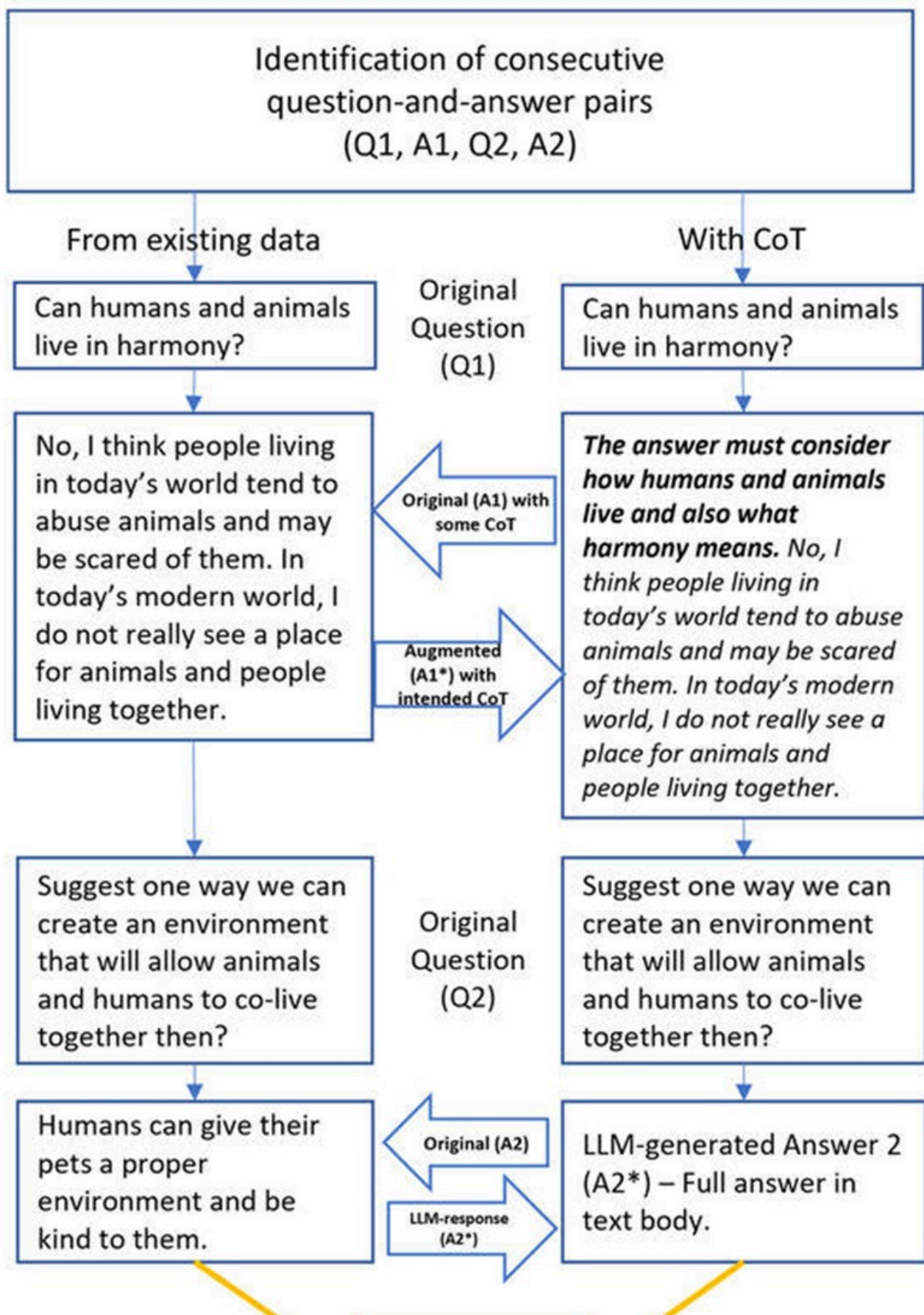
Results and Discussion

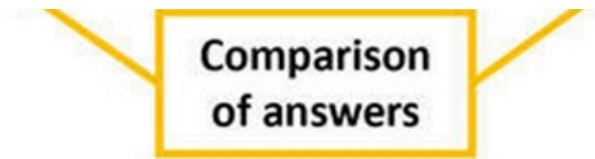
A total of 721 individual conversational contributions were systematically examined, revealing that 272 question–answer pairs contained conceptual matter explicitly aligned with the sustainability dialogue. This indicates that 79% of the student-generated forum submissions maintained direct thematic relevance, while the remaining conversational contributions were categorized as trivial entries or casual socializing and were consequently excluded from deeper analysis. Within the subset of relevant question–answer pairings, researchers mapped 149 distinct student discussion tracks exhibiting a minimum structural depth of 3 steps, with the most extended conversation reaching a depth of 9 steps and involving three separate student contributors.

Two representative conversational tracks containing complete question–answer pairs are detailed below to demonstrate how CoT approaches can be deployed to nurture progressive concepts within knowledge-generative and collaborative frameworks. Both case examples display the sequential progression of existing data harvested from student discourse, tracking how these responses can be augmented to generate a more comprehensive and conceptually flexible set of answers via the LLM. The ultimate outputs are weighed against each other, with detailed theoretical evaluations. **Figure 2** illustrates the initial case example, where an augmented input configured with CoT mechanisms (A1*) is funneled into the LLM to yield an optimized response iteration (A2*). **Figure 3** outlines the secondary case example, wherein the original human response displays an inherent, rudimentary semblance of CoT logic, yet was still augmented using the structured CoT methodology; the subsequent outcomes highlight the compounding value of the CoT model in accelerated optimization of existing concepts and sustaining prolonged conceptual development.









First illustration from empirical data

The AI-synthesized output (A2*) represents a more exhaustive commentary containing an expanded degree of contextual detail relative to the baseline human reply (A2), and its full transcript is paraphrased academically below:

An primary benefit of adopting meat substitutes derived from plants centers on their capacity to significantly minimize ecological degradation relative to conventional livestock farming. The manufacturing of plant-formulated proteins typically demands a smaller allocation of natural resources and yields substantially lower volumes of greenhouse gas emissions than the breeding, rearing, and slaughtering of livestock. Consequently, shifting to these substitutes can play a vital role in shrinking the collective carbon footprint generated by modern agricultural sectors. Furthermore, contemporary plant-based meat variants successfully replicate the flavor profiles and physical textures of conventional animal products, offering a practical alternative for consumers aiming to scale back their meat consumption due to medical, ethical, or environmental concerns without forcing them to completely abandon the familiar culinary styles and flavor profiles they traditionally enjoy.

Second illustration from empirical data

The text synthesized by the AI application (A2*) similarly exhibits greater length and structural complexity compared to the human-authored input (A2); this machine-generated response is reproduced below in its entirety:

Without a doubt, establishing a balanced framework where human populations and wildlife can peacefully coexist is achievable, though it demands deliberate intent and mutual accommodation. One practical pathway to execute this involves advancing animal preservation initiatives and revitalizing degraded wilderness zones: through the safeguarding and rehabilitation of organic ecosystems, communities can secure the territorial expanses and natural provisions wildlife requires to flourish. This strategy centers on conserving pristine environments, cultivating indigenous plant life, and establishing biological migration pathways that enable wild populations to travel unhindered across fragmented habitats. When ecological balances are properly maintained, they possess the capacity to accommodate a wide array of species, human populations included, while simultaneously decreasing territorial friction. Furthermore, cultivating public awareness regarding the value of biological diversity and the distinct ecological contribution of each organism can cultivate deeper human empathy and systemic understanding. Instructing public cohorts on ethical conduct near wild animals—such as avoiding feeding practices and halting territorial displacement—can markedly curtail adversarial encounters.

Discussion and Limitations

The application of CoT prompting has been investigated as a functional framework to stimulate multi-stage analytical trajectories inside large language models for knowledge-generative targets. Within the specific sphere of common-sense logic—which interfaces directly with knowledge creation and collective knowledge-building methodologies—the recorded outcomes indicate that the primary utility extends beyond simply hitting basic logical goals. Instead, this mechanism yields a marked optimization in the structural standard of AI-synthesized material and core concepts, introducing a more heterogenous array of perspectives that empowers learners to elevate their personal work. This dynamic is clearly manifested across our empirical results, and the subsequent analytical breakdown explicates the structural variations separating the two illustrative cases.

In the initial case study, the human-authored entry supplied by a student functions as a straightforward, linear response that serves primarily to initiate an ensuing peer dialogue. Nonetheless, it lacks the contextual depth and thematic complexity required to stimulate the rigorous critical evaluation that would enable other community members to refine their conceptual frameworks. Conversely, the output derived from the LLM constitutes not merely an elaborated and multi-

faceted explanation, but one that actively harnessed embedded CoT exemplars to direct its processing path, thereby detailing the specific underlying variables explaining why plant-based proteins yield minimized greenhouse gas outputs. The language framework also generated an unprompted secondary outcome along with supplementary explanations; though not strictly requested by the input parameters, this appears to be an intrinsic stylistic feature that the language architecture selected for integration within its output stream.

In the secondary case study, when the baseline human response (A1) was augmented with supplementary exemplars built around a CoT prompt architecture, the resulting LLM output highlighted the profound qualitative shift that contextual CoT models introduce to guide the system toward outputting a significantly more diverse and high-caliber response. Beyond merely supplying expanded descriptions to address the core problem, the synthesized text assumed a predictive and even instructional tone, offering explicit recommendations regarding how public education programs centered on ethical wildlife interaction can actively mitigate negative interspecies friction.

The present investigation has isolated merely two specific illustrative threads out of the broader pool of valid peer interactions logged by students over the course of the research window. What these case examples clarify is that the linguistic foundation of CoT analytical pathways renders them universally applicable across diverse contexts, matching the fundamental behavioral findings documented by Wei *et al.* [13] regarding the latent logical capacities of generative AI tools. Nonetheless, by embedding CoT prompt design as a core mechanic for synthesizing high-tier answers to refine student-led conceptual work, we are arguably witnessing a calibration of technological advancements to satisfy the evolving structural requirements of learners operating within modern, knowledge-driven societal structures [17]. Guided by this operational strategy, CoT prompting can be advanced beyond standard computational logic tasks to output a broader, more diversified spectrum of intellectual feedback that students would likely struggle to formulate independently of generative AI platforms, thereby serving active knowledge-creation and collective knowledge-building objectives.

Admittedly, certain systemic limitations persist that cannot be completely neutralized under current technological paradigms. It is frequently argued that CoT structures are engineered to mimic the internal cognitive paths of human thinkers; yet, this correlation remains unproven, given that the deep neural networks driving contemporary LLM applications excel at extracting information patterns but may lack the capacity to execute reliable, independent logical reasoning [28]. Ultimately, large language models are engineered by training architectures on massive datasets to output human-resembling text strings matching natural language inputs, meaning there is no structural guarantee of an accurate analytical path, which can ultimately yield a mixture of valid and erroneous conclusions.

Furthermore, although this research draws upon empirical records harvested from authentic student conversations, it is highly improbable that the average student will independently structure their inquiries using CoT exemplars when initiating few-shot prompt sequences. It is entirely feasible, however, to systematically train students on the correct methods for interacting with LLM platforms to ensure the technology is used appropriately and for its intended pedagogical purpose. Even so, this reality implies that utilizing CoT structures to manually reinforce student-generated inquiries represents a highly labor-intensive process that must be executed by hand at this evolutionary stage; successfully scaling few-shot CoT prompting will likely depend on the ongoing refinement of these technical systems, or a conceptual shift toward transforming few-shot methodologies into zero-shot generalizations driven by deep reinforcement learning models. Finally, the execution of CoT logic remains entirely dependent on LLM architectures; consequently, this specific operational framework may currently remain cost-prohibitive or structurally unavailable for real-time educational applications that rely on smaller, localized models due to strict budget constraints and hardware limitations.

Conclusion

This investigation offers a definitive perspective on how student populations and instructional designers can successfully leverage LLMs to reinforce knowledge creation and collaborative knowledge-building frameworks by integrating few-shot exemplars into CoT prompt designs. Although the target prompts were manually configured around authentic student inquiries, the empirical findings confirm that within these genuine instructional environments, CoT prompting models

successfully generate enhanced responses that exhibit significantly deeper insight than the original student inputs, which are frequently constrained in both factual density and conceptual scope.

Broadly speaking, while it is technically possible for individual students to naturally formulate inquiries that mirror the structural design of few-shot CoT prompts to facilitate their personal learning, such behavior is exceedingly rare in practice. Consequently, we argue based on the outcomes of this research that the systematic integration of CoT frameworks within student inquiry models should be established as a core element of contemporary digital literacy and instructional scaffolding, providing the structural support necessary to formulate the highly critical questions required to drive active knowledge creation and community knowledge building.

Acknowledgements

The views expressed in this paper are those of the authors and do not necessarily represent the views of the host institution. The research team would like to thank the instructors and student participants involved in this study.

Conflict of interest

None

Financial support

None

Ethics statement

The study was conducted in accordance with the Personal Data Protection Act, and approved by the Institutional Review Board of Nanyang Technological University (IRB-2019-10-034, approved October 2019).

Received: 07 Dec 2023

Revised: 25 Jan 2024

Accepted: 24 Mar 2024

Published online: 15 September 2024

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

Hadi MU, Al Tashi Q, Qureshi R, Shah A, Muneer A, Irfan M, et al. A survey on large language models: applications, challenges, limitations, and practical usage. *TechRxiv*. 2023.
<https://doi.org/10.36227/techrxiv.23589741.v1>.

Paavola S, Lipponen L, Hakkarainen K. Models of innovative knowledge communities and three metaphors of learning. *Rev Educ Res*. 2004;74(4):557-76.

<https://doi.org/10.3102/00346543074004557>.

Bereiter C, Scardamalia M. Learning to work creatively with knowledge. In: De Corte E, Verschaffel L, Entwistle N, van Merriënboer J, editors. *Powerful learning environments: unraveling basic components and dimensions*. Oxford: Pergamon Press; 2003. p.55-68.

Scardamalia M. CSILE/Knowledge Forum. In: *Education and technology: an encyclopedia*. Santa Barbara: ABC-CLIO; 2004. p.183-92.

Hakkarainen K, Paavola S, Kangas K, Seitamaa-Hakkarainen P. Social perspectives on collaborative learning: toward collaborative knowledge creation. In: Hmelo-Silver CE, Chinn CA, Chan C, O'Donnell A, editors. *International handbook of collaborative learning*. New York: Routledge; 2013. p.57-73.

Lee AVY, Tan SC. Promising ideas for collective advancement of communal knowledge using temporal analytics and cluster analysis. *J Learn Anal*. 2017;4(3):76-101.
<https://doi.org/10.18608/jla.2017.43.5>.

Liu V, Chilton LB. Design guidelines for prompt engineering text-to-image generative models. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*; 2022 Apr 29-May 5; New Orleans, LA, USA. New York: ACM; 2022. p.1-23.
<https://doi.org/10.1145/3491102.3501825>.

Rae JW, Borgeaud S, Cai T, Millican K, Hoffmann J, Song F, et al. Scaling language models: methods, analysis and insights from training Gopher. *arXiv*. 2021;arXiv:2112.11446.

Cobbe K, Kosaraju V, Bavarian M, Chen M, Jun H, Kaiser L, et al. Training verifiers to solve math word problems. *arXiv*. 2021;arXiv:2110.14168.

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;33:1877-901.

White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, et al. A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv*. 2023;arXiv:2302.11382.

Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput Surv*. 2023;55:1-35.

Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv Neural Inf Process Syst*. 2022;35:24824-37.

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. *Commun ACM*. 2020;63:139-44.

Westerlund M. The emergence of deepfake technology: a review. *Technol Innov Manag Rev*. 2019;9:39-52.

Opara E, Mfon-Ette Theresa A, Aduke TC. ChatGPT for teaching, learning and research: prospects and challenges. *Glob Acad J Humanit Soc Sci*. 2023;5:33-40.

Lee AVY. Staying ahead with generative artificial intelligence for learning: navigating challenges and opportunities with 5Ts and 3Rs. *Asia Pac J Educ*. 2024;44(1):81-93.
<https://doi.org/10.1080/02188791.2024.2305171>.

Lee AVY, Tan SC, Teo CL. Designs and practices using generative AI for sustainable student discourse and knowledge creation. *Smart Learn Environ*. 2023;10:59.
<https://doi.org/10.1186/s40561-023-00279-1>.

Halaweh M. ChatGPT in education: strategies for responsible implementation. *Contemp Educ Technol*. 2023;15:ep421.

Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. arXiv. 2021;arXiv:2104.08691.

Zhang Z, Zhang A, Li M, Smola A. Automatic chain of thought prompting in large language models. arXiv. 2022;arXiv:2210.03493.

Shum K, Diao S, Zhang T. Automatic prompt augmentation and selection with chain-of-thought from labeled data. arXiv. 2023;arXiv:2302.12822.

Talmor A, Herzig J, Lourie N, Berant J. CommonsenseQA: a question answering challenge targeting commonsense knowledge. arXiv. 2018;arXiv:1811.00937.

Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S. Self-consistency improves chain of thought reasoning in language models. arXiv. 2022;arXiv:2203.11171.

Evans J. Intuition and reasoning: a dual-process perspective. *Psychol Inq*. 2010;21:313-26.

Tan SC, Chan C, Bielaczyc K, Ma L, Scardamalia M, Bereiter C. Knowledge building: aligning education with needs for knowledge creation in the digital age. *Educ Technol Res Dev*. 2021;69(4):2243-66.
<https://doi.org/10.1007/s11423-020-09914-x>.

Scardamalia M. Collective cognitive responsibility for the advancement of knowledge. In: *Liberal education in a knowledge society*. Chicago: Open Court; 2002. p.67-98.

Liu X, Wu Z, Wu X, Lu P, Chang KW, Feng Y. Are LLMs capable of data-based statistical and causal reasoning? Benchmarking advanced quantitative reasoning with data. arXiv. 2024;arXiv:2402.17644.